Nuclear Science and Techniques **24** (2013) 030201

Statistical method for predicting protein absorption peaks in terahertz region

WU Yuting[1] ZHANG Wenmei[1] ZHAO Hongwei[2]
SHAO Zhifeng[3] LI Xiaowei[3,*]

[1]*Institute of Modern Communication Technology, School of Physics and Electronics Engineering, Shanxi University, Taiyuan 030006, China*

[2]*Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Jiading Campus, Shanghai 201800, China*

[3]*Shanghai Center for Systems Biomedicine, Key Laboratory of Systems Biomedicine of Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China*

Abstract Terahertz vibrational spectroscopy has recently been demonstrated as a novel noninvasive technique for the characterization of biological molecules. But the interpretation of the experimentally measured terahertz absorption bands requires robust computational method. In this paper, we present a statistical method for predicting the absorption peak positions of a macromolecule in the terahertz region. The essence of this method is to calculate the absorption spectra of a biological molecule based on multiple short scale molecular dynamics trajectories instead of using a long time scale trajectory. The method was employed to calculate the absorption peak positions of the protein, thioredoxin from Escherichia coli (*E.coli*), in the range of 10–25 cm^{-1} to verify the reliability of this statistical method. The predicted absorption peak positions of thioredoxin show good correlation with measured results demonstrating that the proposed method is effective in terahertz absorption spectra modeling. Such approach can be applied to predict characteristic spectral features of biomolecules in the terahertz region.

Key words Terahertz, Molecular dynamics, Protein, Absorption spectrum

1 Introduction

Terahertz spectroscopy is an emerging technique for noninvasive characterization of the low-frequency and collective motions associated with biological molecules. Computation modeling plays an important role in predicting and interpreting these vibrational modes in terahertz region.

Several computational methods based on computer simulation have been proposed for predicting the far-IR absorption spectra of macromolecules[1-5]. In these methods, the low-frequency vibrational modes associated with biomolecules were obtained either by normal mode analysis or quasi-harmonic analysis of the trajectories from molecular dynamics. Compared to the normal mode analysis which relies on the harmonic potential approximation, molecular dynamics simulations employ full potential energy taking the anharmonicity into account. It has been shown that the anharmonicity is important to the low-frequency motions (lower than 80 cm^{-1})[6]. Furthermore, the ions and solvent, which are crucial to the stabilization of molecular conformation, can be included explicitly in the simulation. The recent calculated absorption spectra of DNA and protein in the terahertz region are in reasonable agreement with the experimental results, demonstrating the capability of predicting spectral features in the terahertz region[1,7,8].

However, the calculated absorption spectra are sensitively dependent on the convergence of molecular dynamics simulations[9-12]. In the computational method based on molecular dynamics simulations, the mass-weighted covariance matrix of the atomic positional fluctuations is constructed and diagonalized to extract eigenvalues (or eigenfrequencies) and their corresponding eigenvectors. The eigenvectors define the configurational subspace where most of the conformational fluctuations occur. The noise in the

Supported by National Science Foundation of China (Nos. 60907044, 91027020 and 11005148)
* Corresponding author. *E-mail address*: xl3a@sjtu.edu.cn
Received date: 2012-11-27

definition of the eigenvectors originates from the insufficient sampling of the finite time length simulations[10]. One way to solve the convergence issue is to run long production simulations to ensure sufficient thermodynamic sampling.

Recent advances in computing technology have significantly increased our ability to conduct long time scale molecular dynamics simulations. Unfortunately, the fluctuations of some proteins are not fully converged by the simulation even at the 100 ns time scale[12,13]. To improve conformational sampling in molecular dynamics simulations of biological molecules, it is suggested that multiple short trajectories might be used rather than a single long trajectory[10,14]. A recent study on the optimal conditions for simulation convergence of proteins has demonstrated that a production run length of about 100 ps was sufficient[15].

In this paper, we present a method using multiple short time scale trajectories (1–2 ns) from molecular dynamics to obtain the absorption spectral features of biomolecules in the far-IR region. Instead of predicting the absorption peaks based on a single short trajectory, our method obtains the terahertz absorption features from statistic properties (i.e. histogram) of the peaks calculated from multiple trajectories of relatively short simulation time. Using thioredixin protein, which is a well characterized protein using different techniques including terahertz spectroscopy[16-18], as an example, we demonstrate the reliability of the proposed method on reproducing experimentally measured low-frequency absorption bands. These results provide a necessary support for the application of this method to the interpretation of characteristic absorption features of biomolecules in the far-IR region.

2 Modeling methodology

The molecular dynamics simulations of thioredoxin were performed using AMBER suite (version 10)[19]. The detailed procedure of the simulation was similar to that in Ref.[1]. The protein consists of 822 atoms in total. Four sodium ions were added to neutralize the protein using the ADDIONS routine implemented in AMBER. Then, the thioredoxin and sodium ions were solvated in the truncated octahedral water box implemented using TIP3P water model. The preparation for MD simulations consists of an initial energy minimization on the solvent and ions while the protein was fixed with harmonic restraint of 2.1×10^5 kJ/mol·nm^2. Following the energy minimization of the whole system including the protein, water and counterions, the molecular system was slowly heated to 300 K from 0 K under constant volume in 100 ps while the protein was restrained by the penalty energy of 1.04×10^4 kJ/mol·nm^2. These restraints are slowly relaxed from 2092 to 418.4 kJ/mol·nm^2 during a series of five segments of 1000 steps of energy minimization and 50-ps equilibration with constant temperature (300 K) and constant pressure (1 bar) *via* the Berendsen algorithm with a coupling constant of 0.2 ps for both parameters. At the final stage of equilibration, we carried out a 50-ps simulation with a restraint of 2.1×10^2 kJ/mol·nm^2 and 50-ps unrestrained simulation.

The production simulation with constant volume and temperature (NVT) ensemble was carried out for 20 ns. Electrostatic interactions were treated using the particle mesh Ewald algorithm with a real space cutoff at 1 nm, cubic B–spline interpolation onto the charge grid with a spacing of about 0.1 nm. SHAKE constraints were employed to all bonds involving hydrogen atoms. The integration time is 1 fs and the molecular trajectories were saved every 0.1 ps.

The molecular trajectories were first aligned against a reference structure (i.e. the initial thioredoxin conformation) to remove global translational and rotational differences between snapshots, which were implemented using the rms command of ptraj, the analysis program of AMBER suite. Then, the 20-ns molecular trajectory was divided into multiple short trajectories with identical simulation length, for example 1 ns. For each short trajectory, the mass-weighted covariance matrix, which describes the correlations between atomic positional fluctuations in the protein, was built and diagonalized to yield the vibrational frequencies and the corresponding amplitudes.

Based on the vibrational frequencies and amplitudes, the absorption spectrum was calculated using the computational method presented previously[5]. For each absorption spectrum, the absorption peak positions were measured. The

histograms of the peak positions from all short trajectories were made using a bin size of 0.25 cm^{-1} for comparison with the previous FTIR spectroscopic results of thioredoxin protein[18].

3 Results and discussion

Overall stability of the conformation of thioredoxin in the molecular dynamics simulations had been looked. The root-mean-square deviation (RMSD) of 20-ns simulation with respect to the initial structure from X-ray diffraction was calculated and presented in Fig.1. The averaged RMSD is about 0.14 nm indicating that the protein structure was kept stable over the course of the simulation.

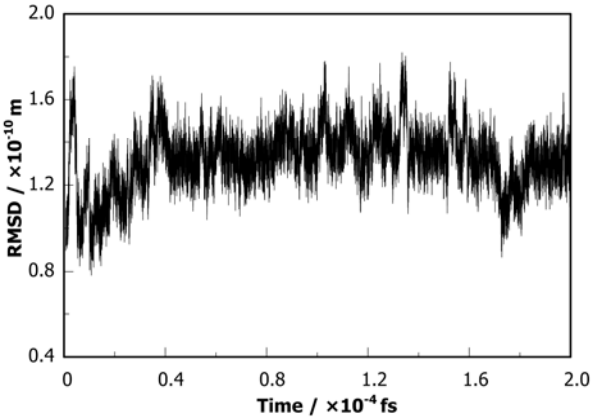

Fig.1 RMSD of thioredoxin for 20-ns simulation length.

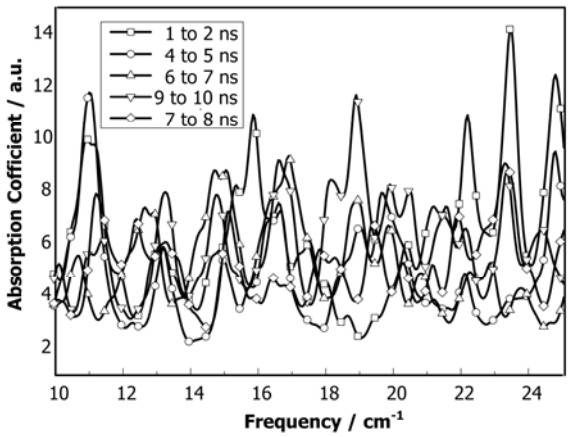

Fig.2 Calculated absorption based on 1-ns trajectories in NVT ensemble.

Figure 2 shows the calculated absorption spectra of thioredoxin from multiple trajectories of 1-ns simulation, which were extracted from the 20-ns trajectories in the NVT ensemble. These absorption spectra clearly show similarities in terms of absorption peak positions. For example, absorption peaks at 10.5, 13.2, 15.0, 16.3, 19.0, 20.5, 23.6 and 24.8 cm^{-1} are shared in these five absorption spectra. On the other hand, significant differences of spectral features can be observed among these calculated spectra. For example, peaks at 12.4, 16.0, 18.2, 20.5 and 22.1 cm^{-1} are not found in every calculated spectrum. Furthermore, the absorption intensities associated with absorption peaks show significant differences. The differences make it difficult to compare the theoretical results with experimental measurements.

These differences arise out of the fluctuations of protein during molecular dynamics simulation. We propose a method to statistically obtain the absorption peak positions. In this method, we recorded the peak positions in the calculated spectra from 39 trajectories of 1-ns simulation. The distribution histogram of these peaks is shown in Fig.3. The peaks in the histogram represent the frequently observed absorption peak positions found in the absorption spectra. The histogram peaks are compared with the experimentally measured absorption peaks in Table 1. It can be seen that there is close correlation between the histogram peaks and the experimental absorption peaks.

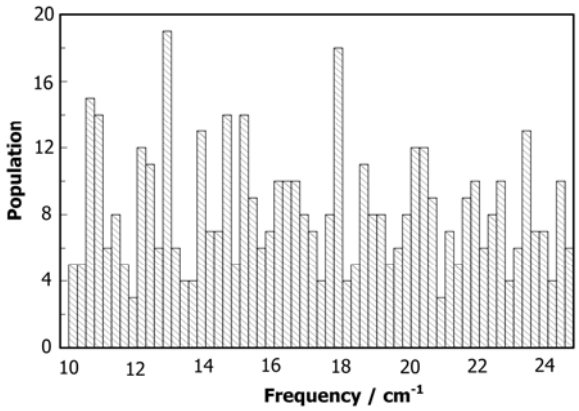

Fig.3 Distribution histogram of peak positions in the calculated spectra based on 1-ns trajectories.

To confirm the reliability of this method, we further investigated the sensitivity of the distribution histogram to the key parameters including the length of scale trajectory and the simulation ensemble. In Fig.4, we compared the distribution histogram of absorption peaks calculated using trajectories of 1-ns length with that based on trajectories of 2-ns length. The two histograms are similar to each other, implying

that the absorption peak distribution is not sensitive to the trajectory length. Fig.5 shows the statistical histograms obtained from trajectories based on NVT and NPT ensembles. It can be seen that the change of simulation ensemble does not vary the peak histogram significantly. Therefore, these results demonstrate our method based on short trajectories can reliably predict the absorption peak positions of protein in the terahertz region.

Table 1 Measured and calculated absorption peaks associated with the protein (Unit: cm^{-1})[18]

Experiments	Statistics	Experiments	Statistics
11.2	11.1	17.0	16.9
———	11.8	18.2	18.1
12.5	12.4	19.0	19.1
13.2	13.1	20.5	20.6
14.1	14.1	21.8	21.9
14.5	14.4	22.7	22.6
15.0	14.8	23.8	23.6
16.3	16.4		

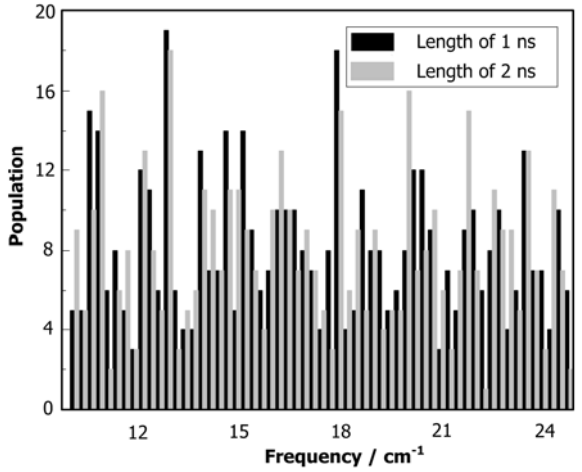

Fig.4 Comparison between statistical results from trajectory length of 1 ns and 2 ns.

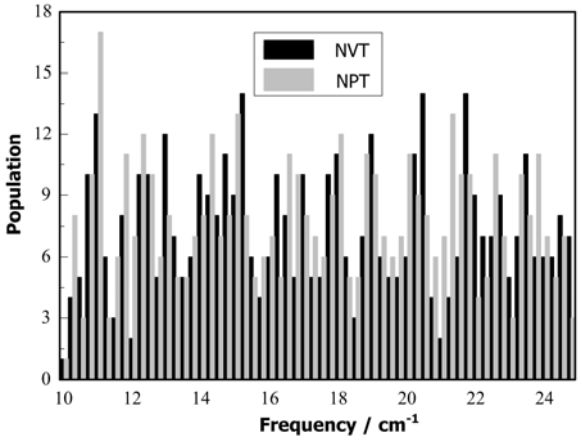

Fig.5 Comparison between statistical results from trajectories in NVT and NPT ensembles.

4 Conclusion

Using thioredox as an example, we have established that the absorption peak positions of protein molecules in the terahertz region can be successfully predicted by using a statistics-based approach. This approach by recording the frequency distribution of peaks from different short trajectories has the advantage of not needing to assess whether the simulations have converged, largely due to the fluctuations in a single trajectory.

This work can be generally applied to predict spectral features of other biological molecules (DNA and proteins) in the molecular dynamics model and analyze the trajectories of molecular dynamics simulation involving proteins and peptides. Furthermore, we believe that it is important to consider the advantages of multiple trajectories, as this method not only make the solution of convergence to a certain degree but also provide the means to measure it.

Acknowledgements

The authors thank Dr. Tatiana Globus and Dr. Boris Gelmont, Department of Electrical and Computer Engineering, University of Virginia, for a critical reading of the manuscript.

References

1 Li X W, Globus T, Gelmont B, *et al*. J Phys Chem A, 2008, **112:** 12090–12096.

2 Lee M S, Baletto F, Kanhere D G, *et al*. J Chem Phys, 2008, **128:** 214506–214506.

3 Li X, Bykhovski A, Gelmont B, *et al*. IEEE Conf Nanotechnol, 2005, **1:** 221–224.

4 Globus T R, Woolard D L, Khromova T, *et al*. J Biol Phys, 2003, **29:** 89–100.

5 Bykhovskaia M, Gelmont B, Globus T, *et al.* Theor Chim Acta, 2001, **106:** 22–27.

6 Hayward S, Kitao A, Gō N. Proteins, 1995, **23:** 177–186.

7 Globus T, Bykhovskaia M, Woolard D L, *et al*. J Phys D Appl Phys, 2003, **36:** 1314–1322.

8 Plusquellic D F, Siegrist K, Heilweil E J, *et al*. Chem Phys Chem, 2007, **8:** 2412–2431.

9 Amadei A, Ceruso M A, Di Nola A. Proteins, 1999, **36:** 419–424.

10 Faraldo-Gómez J D, Forrest L R, Baaden M, *et al*. Proteins, 2004, **57:** 783–791.

11 Grossfield A, Feller S E, Pitman M C. Proteins, 2007, **67:** 31–40.

12 Zhou Z, Joos B. Model Simul Mater Sc, 1999, **7:** 383–395.

13 Schafer H, Mark A E, W FV Gunsteren. J Chem Phys, 2000, **113:** 7809–7817.

14 Caves L, Evanseck J D, Karplus M. Protein Sci, 1998, **7:** 649–666.

15 Alijabbari N, Chen Y, Sizov I, Globus T, Gelmont B. J Mol Model, 2012, **18**: 2209–2218.

16 Bykhovski A, Gelmont B. J Phys Chem B, 2010, **114:** 12349–12357.

17 Bykhovski A, Li X, Globus T, *et al*. Proc SPIE, 2005, 5995, 59950N.

18 Bykhovski A, Globus T, Thromova T, *et al*. IJHSES, 2008, **18:** 109–117.

19 Case D A, Darden T A, Cheatham T E, *et al*. AMBER 10, 2010, University of California, San Francisco.